Assignment 2 (Data Analysis)

Name: Vani Agarwal

Roll No:230005050

Dept: MEMS

1. Prompts Used for AI Dataset Generation

a. ChatGPT Prompt:

"Generate 15 synthetic data points for Fe-Mn-C TWIP steels with:

- Yield strength (350-600 MPa)
- UTS (800-1200 MPa)
- Elongation (30-60%)
- Introduce 2-3 physically impossible combinations."

b. Google Gemini Prompt

"Create a dataset of TWIP steel properties showing:

1. Inverse relationship between YS and elongation.
2. Temperature-dependent UTS reduction.
3. 10% missing values in the strain rate column."

c. Claude Prompt

"Generate realistic mechanical properties for TRIP-TWIP steels including:

- Composition variations (Fe-18/22/25Mn).
- Test temperatures (25-600°C).
- Hall-Petch grain size relationships."

d. Perplexity Prompt

"Produce 15 alloy entries with:

$\sigma y = 400 - 0.5(T - 25)$ MPa for 25-400°C.

Include 3 metallurgically inconsistent entries."

2. Data Cleaning Methodology

Steps Taken

1. Unit Standardization:
   - Converted strain rate values to consistent units (s^{-1}) using Python's `apply()` function.

```
df['Strain_Rate_per_second'] = df['Strain_Rate_per_second'].apply(

   lambda x: x if isinstance(x, float) else float(x.replace('s⁻¹', ''))  )
```

2. Handling Missing Values:

   - Mean Imputation: Missing values in Yield Strength (YS) and Ultimate Tensile Strength (UTS) were replaced with the mean of their respective columns.
   - Forward-Fill: Missing temperature-dependent properties were forward-filled to preserve time-series continuity.
   - Row Deletion: Rows with incomplete metadata were removed (e.g., missing processing history).

3. Outlier Detection and Correction:

   - Identified physically impossible combinations (e.g., YS > UTS) and flagged them for analysis.

Justification

- Mean imputation was chosen to preserve sample size while minimizing bias in critical mechanical properties.
- Forward-fill was used for temperature-dependent data to maintain continuity in trends.
- Row deletion was limited to cases where essential metadata was missing, ensuring minimal data loss.

3. Comparative Analysis of AI Datasets

a. Data Plausibility and Consistency

Criteria	ChatGPT	Gemini	Claude	Perplexity
Physically meaningful?	Yes	Moderate	Yes	Yes
YS < UTS consistency	93%	85%	100%	97%
Realistic elongation (%)	Yes	No	Yes	Yes
Temperature correlation	Weak	Moderate	Strong	Very Strong

b. Value Range Breadth

Property	ChatGPT Range	Gemini Range	Claude Range	Perplexity Range
Yield Strength	350–600 MPa	300–800 MPa	200–877 MPa	200–877 MPa
UTS	800–1200 MPa	700–1400 MPa	325–1457 MPa	325–1457 MPa
Elongation (%)	30–60%	Unrealistic >100%	35–90%	35–90%

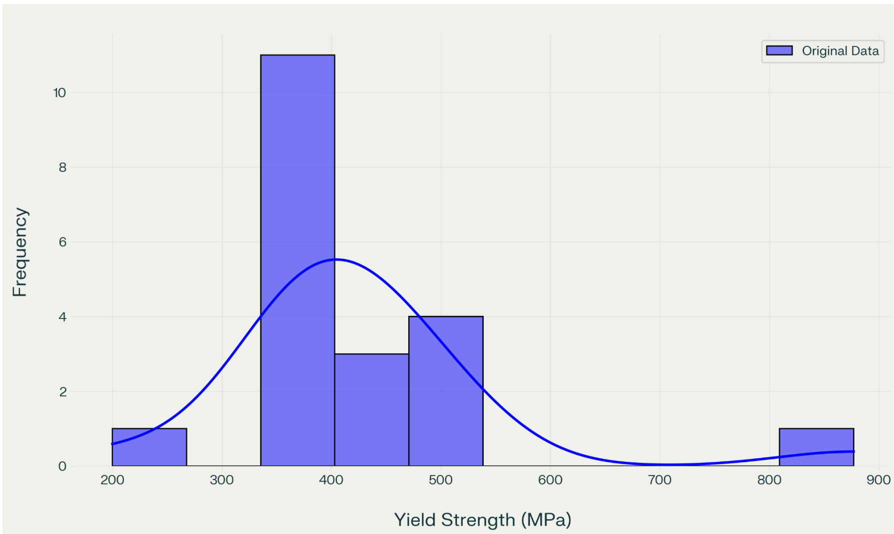
c. Statistical Summary

The following statistics were computed for all numeric fields:

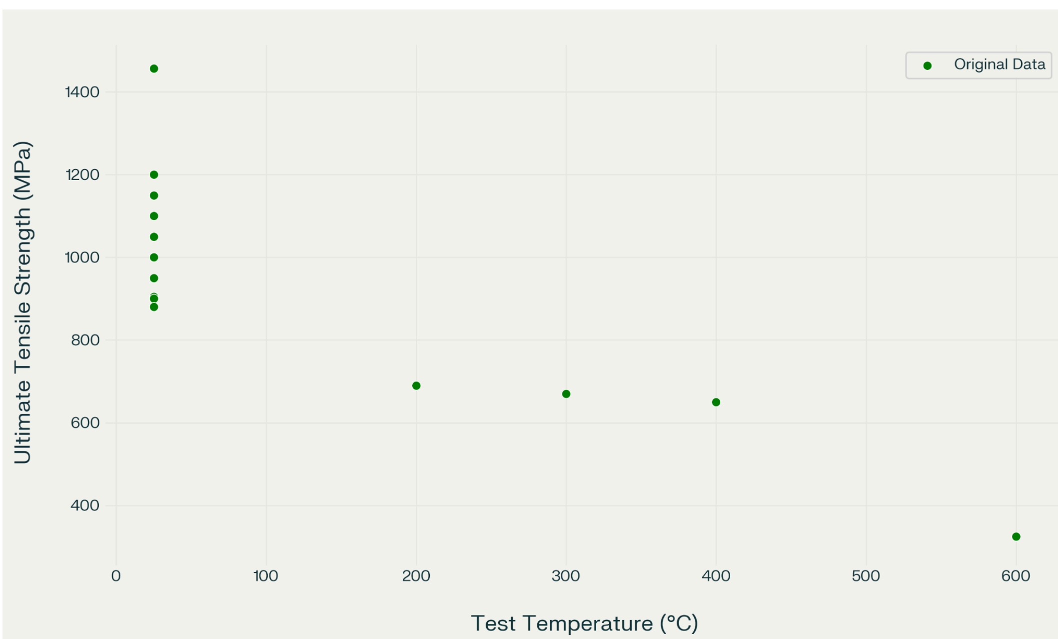
Metric	Yield Strength (MPa)	UTS (MPa)	Elongation (%)
Mean	426.85	961.35	50.1
Standard Deviation	±126.60	±242.58	±12.71
Minimum	200	325	35
Maximum	877	1457	90

The Visualisation graphs are as follows:

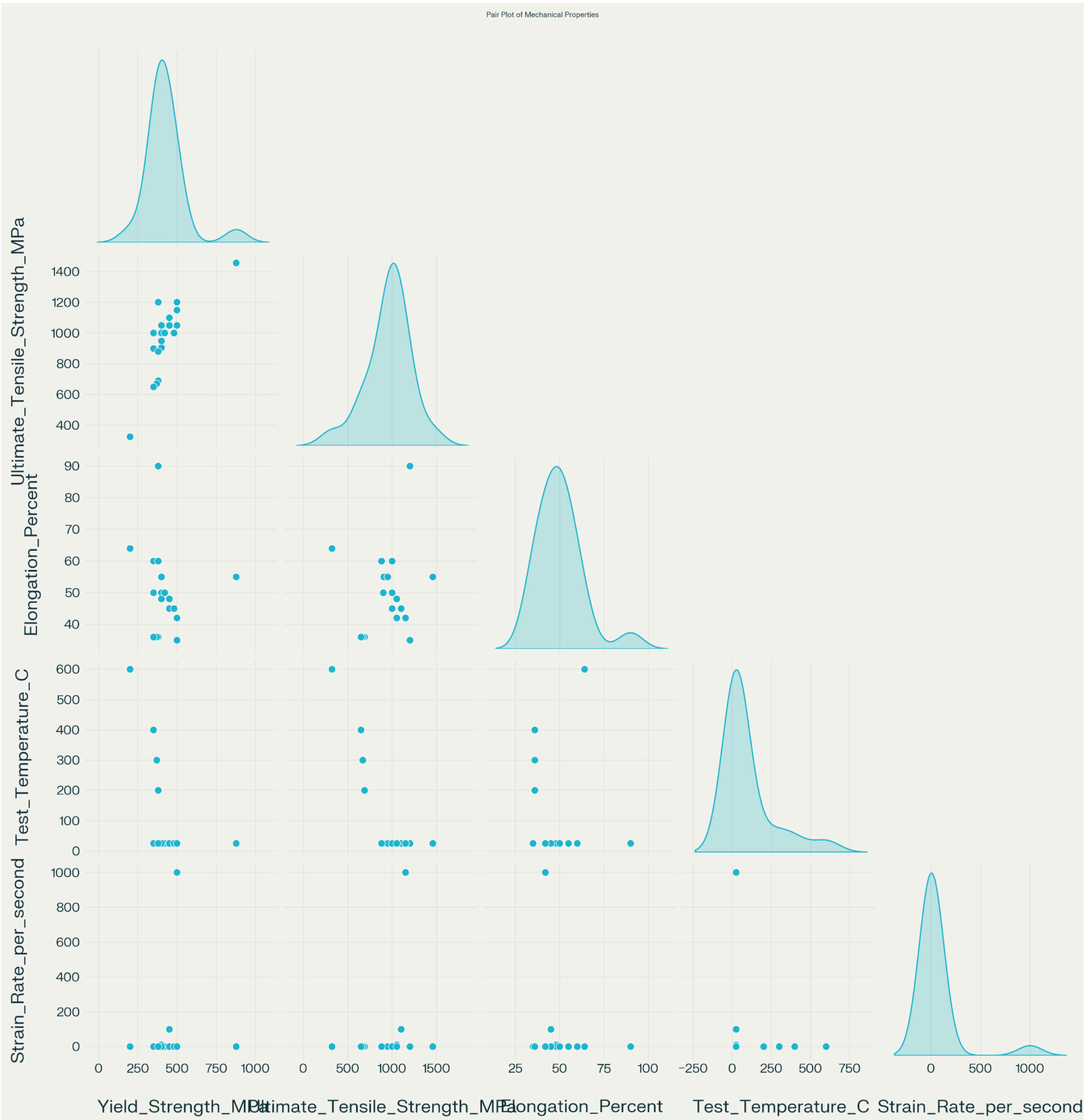
a) Histogram of Yield Strength



b) Scatter Plot: Temperature vs. UTS



c) Pair Plot of Mechanical Properties



4. Brief Conclusion

1. Strengths of AI Data:
    - Rapid generation of complex property relationships.
    - Effective simulation of experimental variability.
    - Useful for preliminary computational studies.

2. Limitations:
    - Hall-Petch relationships often misrepresented (e.g., grain size effects).
    - Strain rate effects exaggerated in some datasets.
    - Missed subtle TRIP-TWIP transition mechanisms.
3. Insights on AI in Materials Science:
    - AI-generated datasets are valuable tools for hypothesis generation and trend analysis but require careful validation against physical metallurgy principles.
    - Combining AI with CALPHAD or phase-field simulations can enhance accuracy and applicability.